BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 3 2007, pages 321–327
doi:10.1093/bioinformatics/btl609

Gene expression

A distribution free summarization method for Affymetrix GeneChip® arrays

Zhongxue Chen[1,2], Monnie McGee[1,*], Qingzhong Liu[3] and Richard H. Scheuermann[2]

[1]Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA, [2]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA and [3]Department of Computer Science, New Mexico Institute of Mining and Technology, Socorro NM 87801, USA

Received on July 18, 2006; revised on November 6, 2006; accepted on November 24, 2006
Advance Access publication December 5, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Affymetrix GeneChip arrays require summarization in order to combine the probe-level intensities into one value representing the expression level of a gene. However, probe intensity measurements are expected to be affected by different levels of non-specific- and cross-hybridization to non-specific transcripts. Here, we present a new summarization technique, the Distribution Free Weighted method (DFW), which uses information about the variability in probe behavior to estimate the extent of non-specific and cross-hybridization for each probe. The contribution of the probe is weighted accordingly during summarization, without making any distributional assumptions for the probe-level data.

Results: We compare DFW with several popular summarization methods on spike-in datasets, via both our own calculations and the 'Affycomp II' competition. The results show that DFW outperforms other methods when sensitivity and specificity are considered simultaneously. With the Affycomp spike-in datasets, the area under the receiver operating characteristic curve for DFW is nearly 1.0 (a perfect value), indicating that DFW can identify all differentially expressed genes with a few false positives. The approach used is also computationally faster than most other methods in current use.

Availability: The R code for DFW is available upon request.

Contact: mmcgee@smu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Using microarrays of polynucleotide probes, researchers can measure the expression levels for tens of thousands genes simultaneously. This provides an opportunity for scientists to investigate the functional relationship between the cellular and physiological processes of biological organisms and their genes at a genome-wide systems level. Of the several types of gene expression microarrays, the Affymetrix GeneChip® is the most widely used.

An Affymetrix GeneChip® can contain from 6000 to >50 000 probe sets that target specific genes, depending on the organism and platform. Usually the number of probe pairs within a probe set is between 11 and 20. For each probe pair, there are two probes. The perfect match (PM) probe is 25 nt in length, and is perfectly complementary to a subsequence of the target mRNA. The mismatch (MM) probe is identical to the corresponding PM probe except that the middle (13th) base is intentionally changed to its Watson–Crick complement. MM probes were originally designed to estimate the background signal of the corresponding PM probes.

The raw microarray intensity data are usually highly 'noisy'. Consequently, before any high level analysis, such as gene selection, classification, or clustering, is executed, a series of preprocessing procedures are usually performed. These preprocessing steps can profoundly affect the results of high-level analyses. A typical preprocessing procedure consists of three steps as follows: background correction, normalization and summarization, not necessarily in this order. The background correction step is typically done in an attempt to remove the part of the raw intensity value that is due to non-specific binding of the labeled target mixture; the normalization step reduces systematic variation in intensity value distributions between chips and the summarization step generates a single expression value for each gene from the probe-level intensity values.

In this paper, we focus on the summarization step. There are several summarization methods in common use. The earliest one, Affymetrix Microarray Suite (MAS 4.0), takes the average of the background corrected intensities of PMs within a probe set after removing the smallest and largest values (AvDiff). MAS 5.0 (Affymetrix, 2002) uses 1-step Tukey Biweight method to get a gene expression summary. Model Based Expression Index (MBEI, Li and Wong 2001a, b) uses a model to estimate the signal based on the original scale. Robust Multi-chip Average (RMA, Bolstad *et al*., 2003; Irizarry *et al*., 2003a, b) obtains gene expression values using the non-parametric and robust median polish approach on the logarithm-transformed probe intensities. However, the RMA background correction method makes the assumption that the probe intensities follow an exponential-normal convolution model. A recently developed summarization procedure, Factor Analysis for Robust Microarray Summarization (FARMS, Hochreiter *et al*., 2006), is also a model-based method that uses logarithm-transformed data.

Model-based methods are heavily dependent on model assumptions, and require estimation of model parameters in order to perform effectively. In practice, some of the assumptions that are

*To whom correspondence should be addressed.

based on distributional characteristics may not be appropriate for microarray data. Furthermore, parameter estimation is not an easy task for microarray data. Maximum likelihood procedures are typically unstable, and EM-based algorithms too slow, due to the large amount of data generated by a typical microarray experiment (Bolstad, 2004; McGee and Chen, 2006).

In addition, few of these models have taken into account non-specific- and cross-hybridization, which can vary among probes within a given probe set. Several observations indicate that non-specific- and cross-hybridization are significant contributors to probe-level intensity values. First, the PM and MM values of a probe set are highly correlated with each other, presumably due to some level of non-specific hybridization of the MM probe to the specific target transcript (M. McGee and Z. Chen, unpublished data). Second, for a substantial fraction of the probe sets (~30% in many datasets), the MM intensity is actually greater than the PM intensity (Irizarry *et al*., 2003b), presumably due to MM cross-hybridization to a different target transcript. Estimating the effects of non-specific- and cross-hybridization is complicated by the fact that the hybridization capabilities for PM probes within a probe set are not the same due to differences in probe content and structure. The difference in probe-level hybridization to the same target transcript has been termed the 'probe effect' (Li and Wong, 2001a, b).

We propose a new non-parametric summarization technique, the Distribution Free Weighted method (DFW). DFW is non-parametric in the sense that no distributional assumptions are made on the probe-level intensities. In its current implementation, no background correction is performed, and quantile normalization, as in RMA and FARMS, is employed for normalization purposes. Also like RMA and FARMS, only the PM probes are used. We compare our new method with MAS 5.0 and its later improved version Probe Logarithmic Intensity Error (PLIER, Affymetrix, 2005), MBEI, RMA, RMA-noBG, Gene Chip RMA (GCRMA, Wu *et al*., 2004) and FARMS. We use two sets of Affymetrix Spike-in data (available at: http://www.affymetrix.com) along with the 'GoldenSpike' dataset (Choe *et al*., 2005), for comparison. The results show that DFW outperforms other methods when both of sensitivity and specificity are considered.

DFW and the method comparisons are implemented in R (Ihaka and Gentleman, 1996) and Bioconductor (Gentleman *et al*., 2004). Both programs are available at http://www.bioconductor.org.

2 METHODS

Since different probes within the same probe set hybridize with different strengths to the same target, a preprocessing method should take these probe effect differences into account. However, for most preprocessing methods, the probe effect for each probe is assumed to be a constant across the probe set (Li and Wong, 2001a, b). It is well accepted that there is a linear relationship between the specific hybridization intensity and the concentration of the target mRNA (Lockhart *et al*., 1996). Indeed, the spike-in datasets are based on this linear relationship assumption for log intensity and log concentration (Lockhart *et al*., 1996). Under this assumption, the relative specific hybridization intensities under different conditions from all PM probes within a probe set should also be consistent. For example, if probe A has stronger hybridization characteristics than probe B, then probe A intensity signal should always be stronger than the probe B signal by some consistent value.

However, we can only estimate the fold differences between experiments based on the observed intensities that contain noise. There are several reasons why the estimated intensity values, even for probe-pairs that are part of the same probe set, are disparate from one another. First, no matter what methods are used, it is difficult to remove all the background noise. The background noise has a large effect on the estimated intensity values, especially when the intensities are low. Second, the amount of non-specific- and cross-hybridization is different from probe to probe. Currently, there are no reasonable methods to remove effects due to non-specific- and cross-hybridization from the observed intensities. Third, some PM probes are not really PM probes for the gene (probe set) they have been assigned to, although they were thought to be so when the chip was designed, due to our evolving knowledge of genome sequence and transcript structures (Dai *et al*., 2005; Harbig *et al*., 2005). Therefore, each PM probe should not be treated equally during summarization.

A good summarization method should not only utilize the information from multiple chips, as RMA, MBEI and FARMS do, but also consider the different qualities of PM probes within a probe set. The new method described here, DFW, is a multi-chip method and takes advantage of information among arrays. The 'hybridization quality' of each PM probe within a probe set is estimated based on the intensity value (log base 2) variability for that probe across all arrays. The final estimated probe set intensities are weighted averages that give larger weights to high-quality PM probes. This approach is based on the postulate that relative intensity values of probes that suffer from non-specific- and cross-hybridization within a probe set will demonstrate higher variability among arrays derived from independent biological samples. The following paragraphs give more detail on the method, and an example of DFW performed on a much-simplified dataset is given in the Supplemental material.

The observed intensities are first logarithm-transformed for each PM probe across arrays. The range (maximum–minimum) of the log intensity value for each PM probe across arrays and the median range value for this probe are calculated. The median-centered range of the log intensity values for PM probe i is denoted by x_i. We denote the maximum absolute value of x_is as Max. We use the Tukey weight function:

$$w_i = \left(1-\left(\frac{x}{\text{Max}}\right)^2\right)^2. \quad (1)$$

And the final weight for probe i is:

$$w_i = \frac{w(x_i)}{\sum_{j=1}^{J} w(x_j)} \quad (2)$$

Where J is the number of PMs within that probe set. If all the ranges are the same, then each PM probe has the same weight. By using this weighting procedure, small weights are given to those PM probes with poor qualities due to unusually high or low variability across arrays. Here, we assess the quality of PM probes across all arrays, as this avoids a common situation where a PM probe may perform well for some arrays or conditions (for example, when the concentrations are high), but has poor behavior for other arrays or conditions.

The summarized expression values (log base 2) of a probe set across arrays are then calculated from the weighted probe intensity values. First, for each probe set, the weighted intensity value (a vector with length corresponding to the number of arrays) is calculated based on the log base 2 PM intensity and the weight from each probe within the probe set. The weighted intensity values are linearly transformed to be between 0 and 1 to give the transformed intensity values. This transformation is necessary in order to standardize the comparison between DEGs and non-DEGs, since we assume that the variability between these two groups is different. More specifically, the differentially expressed genes (DEGs) have larger ranges of intensity values (log base 2) than that of non-DEGs, and this range is much larger than would be expected from the probe effects. Therefore, large variability for a probe across arrays that is still present after the weighting procedure has been completed is evidence of differential expression.

A combination of two measures is used to aide in the detection of DEGs. The weighted range (WR), the range of the weighted intensity values for

Table 1. Average AUC (percent) for Datasets A and B using different summarization methods

Dataset	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIE R
A	100	91	69	60	65	05	26	3
B	100	95	57	65	63	06	40	20

each probe, is calculated as an absolute measure of a probe's variability across experiments. A weighted standard deviation (WSD) is calculated in the same way as the weighted intensity values, where x_i in (1) is replaced by the median-centered standard deviation across arrays. The WSD captures average variability around the mean weighted intensity for each PM probe. The WSD is useful in mitigating the effect of extreme probe intensity values on the WR, thus assuring that large values of their combination point to DEGs, rather than genes containing a few poorly performing probes. Therefore, the WSD of probe intensities across arrays provides additional information.

The final expression value G_i for gene i across arrays is given by

$$ExpValue = \min + c(\mathrm{TIV})(\mathrm{WR})^m(\mathrm{WSD})^n \quad (3)$$

Here, m and n are positive numbers, min is the minimum of the weighted intensity values before the linear transformation, and c is a scale parameter. The parameters m and n are positive, and they are intended to accentuate large differences in the WR and WSD in order to facilitate the selection of DEGs. Default values are set to be $m = 3$, $n = 1$. In practice, m and n larger than 3 results in log expression values that are larger than can be reasonably expected given the detection limits of most scanners. For example, typical values of WR range from 7 to 11 (on a log base 2 scale), while values for WSD range from 5 to 8. Therefore, c is applied globally to make the expression values more manageable. In the analysis of Datasets A and B that follow (explained in the next section), $m = 3$, $n = 1$ and $c = 0.01$. For Dataset C, $m = 3$, $n = 1$ and $c = 1$. More detail on the choice of m and n is given in the discussion section.

3 RESULTS

3.1 Datasets

We compared our new summarization method with others by using three publicly available spike-in datasets (Dataset A–C). Dataset A is the Affymetrix Latin Square spike-in experiment done on the HG-U95Av2 array. For details on this experiment, see the Affymetrix website (http://www.affymetrix.com). Our comparisons used 16 probe sets recognizing spike-in gene transcripts instead of the original 14, following the recommendations of Cope *et al*. (2004). Dataset B is the Affymetrix Latin Square spike-in experiment performed on the HG-U133A array platform. It was originally designed such that 42 probe sets recognize spike-in transcripts (14 spike-in groups with each group containing three probe sets). McGee and Chen (2006) recently found that there are 22 additional probe sets that also recognize spike-in transcripts in Dataset B. Therefore, there are actually 64 spike-in probe sets for Dataset B. Comparisons were done with both the original 42 spike-ins, via the Affycomp II competition, and the 64 new spike-ins.

Dataset C consists of six DrosGenome1 chips (two conditions with three replicates for each) with 3860 probe sets that recognize spike-in transcripts (Choe *et al*., 2005). Among the 3860 probe sets, 1309 recognize transcripts with known fold differences between the two experimental conditions, from 1.2 to 4, and the remaining 2551 probe sets recognizing transcripts included at the same concentration under both conditions.

3.1 Results

The 'Affycomp II' competition (Cope *et al*., 2004; Irizarry *et al*., 2006, http://affycomp.biostat.jhsph.edu) allows comparisons among 54 and 55 (at the time this paper was prepared) public microarray preprocessing methods based on datasets A and B, respectively. The competition uses many comparison statistics, but only the various areas under the Receiver Operating Characteristic (ROC) curve (AUC) statistics are not scale-dependent (Hochreiter *et al*., 2006). Furthermore, the AUC allows comparison based on sensitivity and specificity simultaneously, which are the most important characteristics of a preprocessing method from the standpoint of a researcher. As detailed below, DFW outperforms all other methods with regard to this simultaneous measure of sensitivity and specificity. In addition, DFW achieves a perfect score on three other measures as follows: median SD, null log FC IQR and null log FF 99%, for both datasets A and B.

Weighted AUC values for DFW, RMA, MAS 5.0, MBEI, FARMS and GCRMA, for both Datasets A and B, are given in Table 1. This comparison uses 42 spike-ins for Dataset B. In Tables 4 and 5, we show the results using all 64 spike-ins for Dataset B (McGee and Chen, 2006).

For the various scores related to regression of expression (or concentration) on nominal concentration values, the results given for DFW in the Affycomp II competition are ranked within the bottom third of the list for both Datasets A and B. RMA-noBG has similar rankings on these measures, but its AUC values are not as large as those of DFW.

For the Latin Square datasets, we also compared pairs of experiments that were separated by the same number of permutations (where d = number of permutations), of the Latin Square. Essentially, for $d = 1$, most of the spike-in transcripts differ in concentration by 2-fold (log2 = 1); for $d = 2$, fold difference would be 4 for most of the spike-ins. [See McGee and Chen (2006) for a more complete explanation of d.] Usually, it is harder to detect the true positives for smaller d than for larger d.

The AUC was calculated for a cutoff of various numbers of false positives (e.g. the number of false positives is 5 for Table 2). The values are then standardized so that the area is between 0 and 100%. For Dataset C no cutoffs were set since the fold changes for this dataset are usually very small and many spike-ins cannot be detected as DEGs for a small number of false positives. Instead, we use the number of false positives necessary to obtain all of the true positives to calculate the AUC. The values are then normalized so that the numbers are comparable among methods. Figure 1 is the ROC curve plot of Dataset A for $d = 1$ when the number of false positives is 5. The DFW method detects all of the true positives without including any false positives. For all other methods tested, false positive detection starts to occur well before all true positives are detected. For $d = 2, 3, ..., 7$, similar plots for ROC curves were obtained. In other words, in all situations tested with this dataset, DFW outperformed the other methods using this important performance metric.

Table 2 shows the effects of degree of separation in the Latin Square design on the performance of the various summarization methods as measured by the AUC. For most of the methods, performance improves as the degree of separation increases, showing

Table 2. AUC (percent) comparison for Dataset A (#FP = 5)

d	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5	MBEI	PLIER
1	100.0	69.2	56.6	45.3	58.7	06.3	06.3	04.6
2	100.0	84.9	79.3	76.4	78.4	12.6	20.6	07.9
3	100.0	86.5	86.5	88.2	86.4	18.8	44.8	06.6
4	100.0	93.3	90.2	92.7	93.3	42.9	59.8	06.7
5	100.0	93.6	98.2	99.2	97.5	74.6	70.8	04.0
6	100.0	93.7	99.3	99.6	99.8	84.8	76.3	00.0
7	100.0	93.7	99.6	99.8	99.7	86.2	80.1	00.0

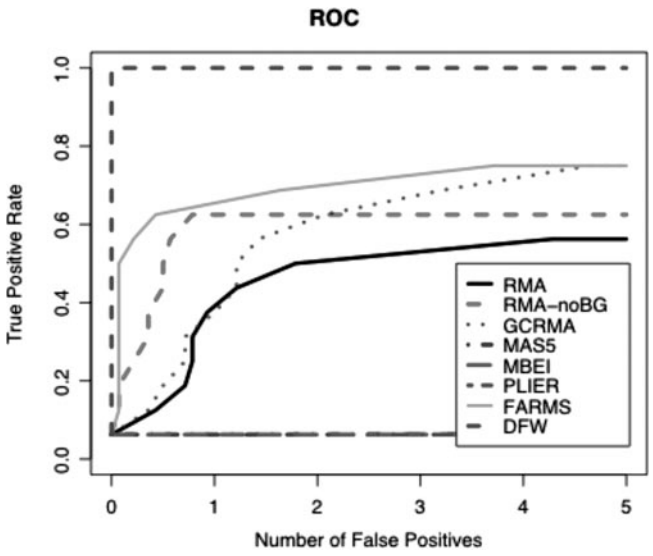

Fig. 1. ROC plots based on Dataset A for DFW and popular preprocessing methods. ROC analysis was performed for $d = 1$ where the maximum number of false positives is 5. DFW detects all the true positives without any false positives.

that as the average fold difference in expression increases, most methods are more easily able to distinguish between real differences in target amounts given the experimental variability. In contrast, DFW is effective at all degrees of separation, even when the difference in target amounts is relatively small (e.g. 2-fold).

AUC analysis is dependent on the range of false positive values considered. Table 3 shows the effect of increasing the extent of false positive values on the AUC results using the various summarization methods. Even at relatively high false negative extents, DFW is found to significantly outperform the other methods using the AUC metric.

Figure 2 shows the ROC curve plot of Dataset B for $d = 1$ when the number of false positives is 10. Once again, the DFW method can obtain all of the true positives with very few false positive (<2 on average) and is superior to the other methods tested. As with Dataset A, similar findings were observed for $d = 2, 3, \ldots, 7$ (Table 4).

Table 5 shows the effect of increasing the extent of false positive values on the AUC results using the various summarization methods for Dataset B. Again, even at relatively high false negative extents, DFW is found to significantly outperform the other methods using the AUC metric.

Table 3. AUC (percent) from Dataset A for given numbers of false positives when $d = 1$

# FP	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5.0	MBEI	PLIER
2	100.0	62.6	36.2	32.0	53.0	6.3	6.3	0.0
5	100.0	69.2	56.6	45.3	58.7	6.3	6.3	4.6
10	100.0	75.4	65.8	52.9	63.8	6.3	10.3	5.4
20	100.0	81.9	72.9	60.6	66.3	6.3	25.3	5.8
40	100.0	87.2	78.8	64.6	67.5	6.3	41.6	6.0

In order to determine if these findings were peculiar to these spike-in datasets with small numbers of true positives, a similar study was performed on the GoldenSpike dataset (Dataset C), in which 1309 probe sets recognize spike-in transcripts with known fold differences (from 1.2 to 4) between the two experimental conditions.

It is important to mention that the use of the Choe 'GoldenSpike' dataset for method comparisons is somewhat controversial. Two recent papers have criticized this dataset (Dabney and Storey, 2006; Irizarry, *et al*., 2006). The former is a reanalysis of the Choe dataset, explaining that the original analysis in the Choe *et al*. paper was not appropriate. In particular, the original analysis did not take into account the fact that the replicated arrays were technical replicates rather than biological replicates, and so the analysis presented here for this dataset will not have included the effects of biological variability in the results.

Irizarry *et al*. (2006) show evidence for an experimental artifact pertaining to the different behavior of features spiked-in to be at a 1:1 ratio between the control and the spike-in experiment. We conjecture that this artifact may be present due to the necessity of adding additional cDNA (mRNA) material as the concentrations were increased. If Choe *et al*. had tried to keep the amount of cDNA constant, it is likely that other artifacts would have been produced in the data. Essentially, in an experiment such as this, it would be very difficult to get around all potential artifacts. Also, we have no way of knowing whether something similar to this might occur in real data. We have included an analysis using the Choe dataset mainly because its large number of differentially expressed transcripts allows for better estimation of false and true positive rates. Furthermore, inclusion of this dataset gives a comparison on data that is quite different in design from the Affymetrix spike-in datasets.

For comparison using Dataset C, we added another method that was advised by Choe *et al*., as a result of their experiments with various combinations of methods for background correction, normalization and summarization. We call this the 'Choe Preferred' (CP) method. This method first corrects the background using the local procedure of MAS 5.0. Normalization is accomplished using quantile normalization, followed by median polish summarization. There is a second normalization after summarization, for which the LOESS procedure is used.

Table 6 shows the AUC calculations using the various summarization methods on Dataset C. The first row is the values of fold difference in transcript levels (spike-in group versus control group)

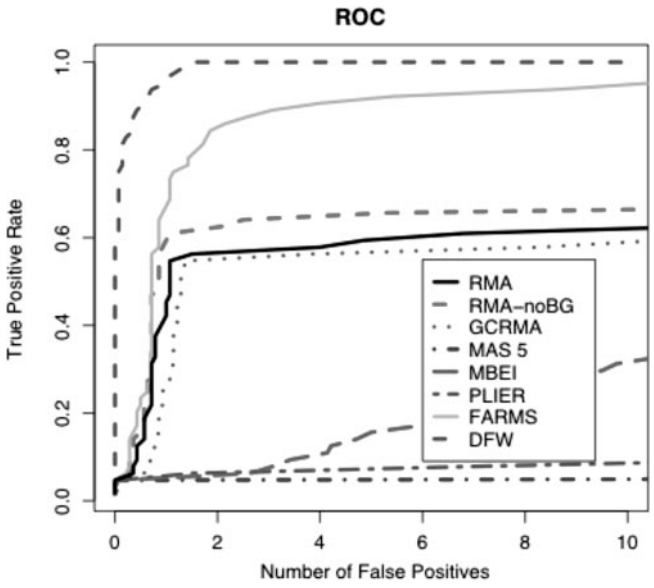

Fig. 2. ROC plots based on Dataset B (with 64 spike-ins) when $d = 1$ and number of false positives is 10. DFW detects all the true positives with ~2 false positives.

Table 4. AUC (percent) comparison for Dataset B (#FP = 10)

d	DFW	FARMS	GCRMA	RMA	RMA-noBG	MAS 5.0	MBEI	PLIER
1	98.6	84.1	51.4	55.0	60.7	4.7	15.5	7.0
2	98.7	86.4	58.9	67.4	63.5	10.6	36.3	19.7
3	99.3	93.2	65.8	80.6	81.5	27.6	48.4	43.4
4	99.7	94.7	78.3	88.4	91.5	51.4	64.9	65.8
5	99.9	97.0	93.4	94.8	96.1	76.0	76.2	82.2
6	100.0	97.8	95.5	96.5	96.9	98.5	81.0	88.7
7	100.0	98.2	97.6	98.2	98.3	90.7	83.6	91.2

and the second row gives the number of spike-in probe sets that have the corresponding fold differences given in row 1. Some of the spike-ins do not have the exact fold changes listed. For example, there are 167 out of 172 spike-ins that have exact fold change of 1.2 but 5 of the 172 have fold differences >1 but <1.2. Choe *et al*. (2005) explain this phenomenon in their paper. The values in the parenthesis of row 2 are the numbers of spike-ins that have the exact fold differences indicated in column 1. The first two largest values of AUC for each fold difference category of spike-ins are highlighted in bold in Table 6. DFW and GCRMA almost always give the highest AUC values. Usually GCRMA has slightly larger values of the AUC than does DFW.

Dataset C was created to mimic real data as closely as possible; therefore, it is noisier than the Latin Square datasets, and likely requires more background correction and normalization. In an attempt to improve the performance of DFW on the Choe dataset, we used the exponential-normal background correction (RMA) and zonal correction (MAS 5.0) in concert with DFW. We further investigated the effect of these background correction methods in combination with various normalization methods, specifically, constant normalization, LOESS normalization and invariant set normalization. The use of the two background methods above improved the performance of DFW, but not above the level of GCRMA. Combinations of various background correction and normalization algorithms also did not improve the performance of DFW above the level of GCRMA. However, the use of LOESS normalization, with no background correction and DFW summarization, had a better performance than GCRMA (AUC 0.96 and 0.95, respectively, for spike-ins with fold change >2.0).

Table 5. AUC (percent) from Dataset B for given numbers of false positives when $d = 1$

# FP	DFW	FARMS	GC-RMA	RMA	RMA-noBG	MAS 5.0	MBEI	PLIER
5	97.1	74.9	45.3	49.3	55.5	4.7	7.9	6.0
10	98.6	84.1	51.4	55.0	60.7	4.7	15.5	7.0
20	99.3	90.1	55.4	60.0	64.0	4.7	28.6	8.5
50	99.7	95.0	58.9	65.7	67.5	4.7	45.1	15.4
100	99.9	97.6	61.3	70.2	69.8	5.7	53.6	24.8

In our analysis, CP performs better than PLIER and sometimes MAS 5.0, but much worse than the others. The discrepancy between its performance in the Choe paper versus ours can be explained by the way in which DEGs were determined in Choe's study. The Choe *et al*. paper used a modified *t*-statistic (Welcome to Cyber-T http://visitor.ics.uci.edu/genex/cybert) to identify DEGs. They based their evaluation of the methods on the number of genes in the top 1000 significant genes that were true spike-ins. For this paper, we used a fold difference based method, as is typically done for Datasets A and B. Fold difference makes more sense for Choe data, since there are a small number of replicates (three for each condition). Hence, the estimated standard deviations, necessary for Cyber T, are not very reliable.

Figures 3 and 4 show the ROC curves for all the methods for Dataset C. In Figure 3, we show curves for all spike-ins with nominal fold change levels equal to 1.2. Figure 4 shows the comparisons for all spike-ins considered simultaneously.

The spike-in datasets are derived from highly specialized experiments, and it is not clear that these results, as good as they may be, will hold for real data (Allison, 2006). To remedy this situation, we compared the performance of DFW to RMA, MAS 5.0, PLIER, FARMS, and MBEI on simulated data. The model used for DEGs in the simulations is given by

$$\text{PM} = \text{pe} + \text{signal} + \text{noise} + \text{fc}$$

where pe = probe effect and fc = fold change. Non-DEGs were simulated using the same model, excluding the fold change. We assume that there are 11 probes per probe set and that there are three 'bad' probes within each probe set. This assumption is made for both DEGs and non-DEGs. Details of the parameters used and the results, including ROC curves, are given in the Supplemental material. Again DFW was found to outperform all methods on these simulations, which were designed to simulate the effects of non-specific and cross-hybridization on PM signals. We have now begun a comprehensive analysis of DFW on real microarray data related

Table 6. AUC (percent) comparison for Dataset C

FC	1.2	1.5	1.7	2.0	2.5	3.0	3.5	4.0	All
Spike-ins	172 (167)	182 (169)	181 (179)	146 (139)	192 (182)	97 (93)	184 (184)	177 (177)	1331
DFW	**89.7**	65.8	61.6	**82.7**	**90.1**	**96.2**	**94.8**	**98.6**	**84.1**
FARMS	87.8	**66.0**	**63.6**	78.3	87.8	94.8	94.0	97.5	82.9
GCRMA	86.0	**74.3**	**78.3**	**88.4**	**91.9**	**96.1**	**96.0**	98.5	**88.1**
RMA	86.3	48.8	45.0	74.5	84.7	92.8	93.9	97.7	76.8
RMA-noBG	**88.8**	59.6	56.8	80.0	87.9	94.9	94.4	98.1	81.6
MAS5.0	53.2	30.6	15.2	21.1	30.1	41.9	58.9	62.5	39.0
MBEI	81.5	40.6	45.9	76.0	87.2	94.5	94.4	**98.9**	76.0
PLIER	12.5	9.6	31.1	51.8	60.9	72.0	83.8	86.6	49.7
CP	37.1	23.2	42.7	62.4	70.5	78.0	84.4	87.2	59.5

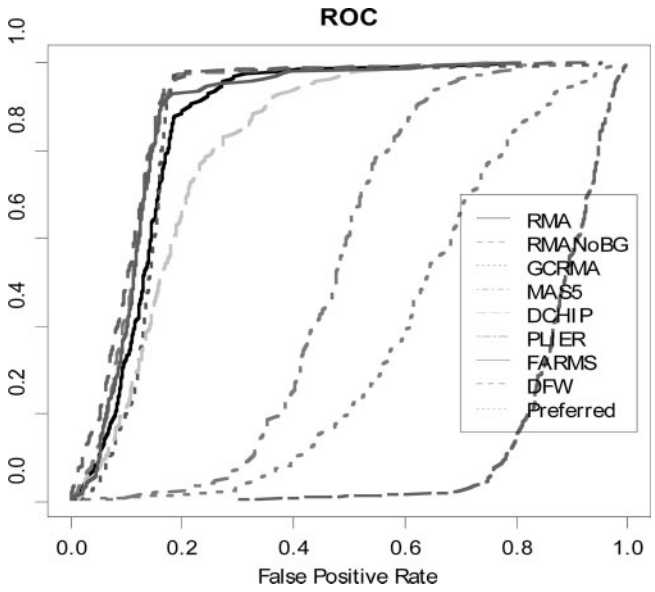

Fig. 3. ROC curves based on Dataset C and when the FC is 1.2. DFW has the largest AUC values.

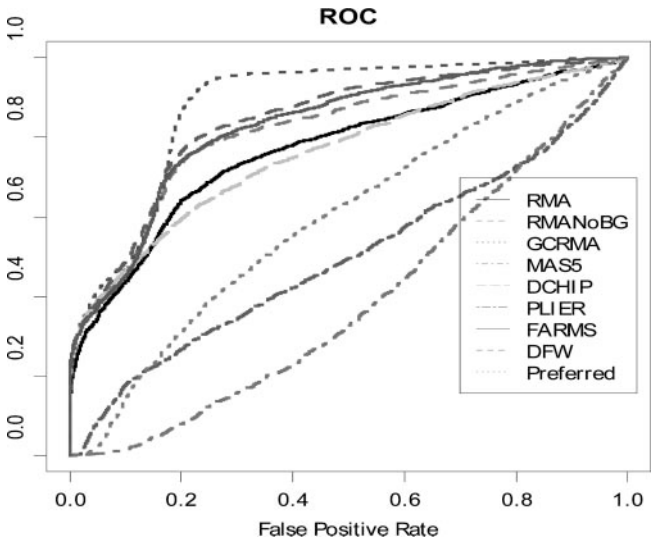

Fig. 4. ROC curves based on Dataset C when all the spike-ins are considered. DFW has the second largest AUC values.

to B lymphocyte transformation, differentiation and function using GeneOntology™ annotation co-clustering and experimental validation approaches (Lee, 2006).

Finally, the computational complexity for each method was estimated by measuring the CPU time, as given on a PowerMac G5 running R Cocoa GUI (Iacus and Urbanek, 2005) with R version 2.2.0 (Table 7). The computational time for DFW is much less than the other methods, such as RMA, MBEI, FARMS and PLIER. This is because DFW is not an iterative method. MBEI and FARMS, for example, require iterative algorithms to estimate the necessary model parameters. In addition, it should be noted that MBEI does not always converge, even for the Latin Square datasets, and is particularly unsuitable for the GoldenSpike dataset due to the small number of arrays.

4 DISCUSSION

DFW outperforms all methods we examined by a large margin, in terms of sensitivity and specificity, on the Affymetrix Latin Square datasets. GCRMA performs slightly better than DFW on Dataset C, in spite of the use of background correction and normalization methods in combination with DFW. The difference in performance can be explained by the fact that the assumptions for the various background correction and normalization methods are violated on the Choe dataset. The exponential-normal and zonal background correction methods both assume that the background intensities are not probe dependent. Due to the differing binding affinities for different probes, this is likely not the case. Furthermore, current normalization methods (including quantile normalization) make the assumption that genes are symmetrically expressed (equally likely to be up or down regulated), an assumption that is clearly violated by the Choe dataset, since all of the spike-in transcripts are designed to be upregulated. Another assumption that is typically implicit in most normalization methods is that the percentage of DEGs is 'small'. To our knowledge, no one has examined the question of 'How small is small?' nor do there exist normalization methods for situations where a large percentage of genes is expected to be differentially expressed. GCRMA does little in the way of normalization, and its background correction method attempts to model the probe-specific effects. Therefore, we suspect that the better

Table 7. CPU time (in seconds) for various methods on the three spike-in datasets (the best times for each method are marked with asterisks)

Method	RMA	RMA-noBG	GCRMA	MAS 5.0	PLIER	MEBI	FARMS	DFW
Dataset A	342	299	214	953	321	869	132	**112***
Dataset B	388	353	210	1064	239	833	1915	**144***
Dataset C	150	147	78	130	**17***	269	280	68

Table 8. AUC values (percent) calculated from method DFW with different m and n

FC	$m3n1$	$m1n1$	$m1n2$	$m1n3$	$m2n1$	$m2n2$	$m3n3$	$m5n5$	$m8n8$
1.2	89.7	89.3	89.3	89.3	89.6	89.6	89.7	89.8	89.8
1.5	65.8	61.9	62.5	62.9	64.3	64.7	66.3	67.6	67.8
1.7	61.6	59.6	60.8	61.5	61.0	61.9	62.7	63.3	62.6
2.0	82.7	81.6	82.5	83.0	82.4	82.9	83.3	83.6	83.7
2.5	90.1	89.8	90.4	90.7	90.3	90.5	90.6	90.7	90.7
3.0	96.2	95.5	95.9	96.2	96.0	96.3	96.5	96.6	96.7
3.5	94.8	94.7	94.9	95.0	94.8	94.9	94.9	94.9	94.8
4.0	98.6	98.5	98.6	98.7	98.6	98.6	98.7	98.7	98.7
All	84.1	82.9	83.4	83.7	83.4	84.0	84.4	84.8	84.7

sensitivity and specificity of GCRMA compared to DFW is due to the unsuitability of current normalization and background correction methods for these data.

The DFW method involves three parameters that must be chosen, namely m, n and c. As mentioned in Section 2, m and n determine how the extra information provided by the range and SD is used. In our experience, larger values of m and n give better results in terms of AUC. However, if m and n are too large, it will be difficult to scale the final expression values to be comparable with the original ones. The parameter c is used to adjust the expression values to a more reasonable scale. In our analyses, the value of c did not effect the AUC calculations.

Table 8 gives the AUC values for different m and n based on Dataset C. Good results were achieved when m and n were between one and three. Similar conclusions were obtained based on Datasets A and B (see Supplementary tables). Therefore, we would expect that $m = 1$ and $n = 3$ would also give good sensitivity and specificity for real data.

5 CONCLUSION

We have proposed a new non-parametric summarization technique, DFW summarization, which uses variability estimates to identify and down-weight probes that may be especially affected by non-specific and cross-hybridization. This method performed well on three different spike-in datasets when sensitivity and specificity are considered simultaneously in ROC/AUC analysis. In addition, DFW requires less computational time compared with other methods. These data suggest that consideration of probe effects related to non-specific and cross-hybridization during the summarization step can significantly improve the results of Affymetrix gene expression microarray preprocessing.

ACKNOWLEDGEMENTS

The authors wish to acknowledge support for this work from Dr Milton Packer, Director, Department of Clinical Sciences, University of Texas Southwestern Medical Center. This research was supported by the National Institutes of Health contracts N01-AI40076 and N01-AI40041 to RHS. Funding to pay the Open Access publication charges for this article was provided by NIH 1N01AI25487.

Conflict of Interest: none declared.

REFERENCES

Affymetrix, Inc. (2002) Statistical algorithms description document.

Affymetrix, Inc. (2005) Technical note: guide to probe logarithmic intensity error (PLIER) estimation.

Allison,D.B. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet*., **7**, 55–65.

Bolstad,B.M. (2004) Low Level Analysis of High-density oligonucleotide array data: Background, normalization and summarization [dissertation]. Department of Statistics, University of California, Berkeley.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligunucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Choe,S.E. *et al*. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control datasets. *Genome Biol*., **6**, R16.1–R16.6.

Cope,L.M. *et al*. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Dabney,A.R. and Storey,J.D. (2006) A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biol*., **7**, 401.

Dai,M. *et al*. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*., **33**, e175.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Harbig,J. *et al*. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*., **33**, e31.

Hochreiter,S. *et al*. (2006) A new summarization method for Affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Iacus,S.M. and Urbanek,S. (2005) , R Cocoa GUI 1.14 (2129), S.M., © R Foundation for Statistical Computing.

Ihaka,R. and Gentleman,R.C. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat*., **5**, 299–314.

Irizarry,R.A. *et al*. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, 1–8.

Irizarry,R.A. *et al*. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Irizarry,R.A. *et al*. (2006) Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.

Lee,J.A. *et al*. (2006) Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. *BMC Bioinformatics*, **7**, 237.

Li,C. and Wong,H.W. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Li,C and Wong,H.W. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol*., **2**, research0032.1–0032.11.

Lockhart,D.J. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675–1680.

McGee,M. and Chen,Z. (2006) New spiked-in probe sets for the Affymetrix HG-U133a Latin square experiment. *COBRA Preprint Series*, No. 5.

Wu,Z. *et al*. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc*., **99**, 909–917.